Estimating Significant Levels in Stepwise Multiple Regression with Randomization Tests

June 2002

Ontario

Ministry of the Environment

Estimating Significant Levels in Stepwise Multiple Regression with Randomization Tests

Martyn N. Futter and Keith M. Somers
Dorest Environmental Science Centre
Ontario Ministry of the Environment
P.O. Box 39
Dorset, ON, P0A 1E0

June 2002

Cette publication technique
n'est disponible qu'en anglais

ISBN 0-7794-4607-0

PIBS 4338e

Executive Summary

Stepwise multiple regression (SMR) is a commonly used model-building tool in environmental biology. Like all parametric methods, the successful use of SMR requires the satisfaction of a set of assumptions including data normality, lack of collinearity among predictor variables, and a sufficiently high ratio of observations (N) to parameters (P). Failure to meet these assumptions can bias tabulated SMR probabilities, and increase the chance for Type I errors. The potential for error is further compounded when variables are transformed or combined and both the transformed and untransformed variables are used in SMR. This practice lowers the ratio of observations to parameters (N:P), affects normality, and potentially increases collinearity among the set of predictors. Herein, we illustrate how randomization methods can be used to provide unbiased estimates of the significance of SMR models using a variety of published environmental data sets.

We use randomization methods to address two questions: (1) what is the likelihood that an observed SMR model could occur due to chance alone if Y is truly random; and (2), how do the randomization-based significance levels differ from the standard tabulated values as predictor collinearity, data normality, and N:P ratio are changed?

Real and simulated data were used to examine the first question. We performed an SMR and recorded the F statistic for the best model. We then randomly re-shuffled the Y variable, re-ran the SMR and recorded the resultant F statistic for each model using the re-shuffled data. The observed F statistic was compared to the set of F statistics from the re-shuffled data to estimate the probability that the observed Y variable was truly random with respect to the predictor variables.

A series of simulated data sets drawn from different underlying frequency distributions, with different levels of collinearity and N:P ratios were generated to address the second question. For each analysis a random "response" vector was generated and an SMR run to predict the Y variable from a simulated predictor matrix. A randomization test was performed by re-shuffling the Y variable, recalculating the sums of squares and associated regression coefficients, and then counting the number of recalculated values that equalled or exceeded the observed values. This approach provided an estimate of the significance of the observed SMR results when Y was truly random with respect to the predictors. The resultant probabilities from the randomized and observed analyses were compared to assess whether all variables in the observed model were significant predictors of the response variable.

Our findings emphasise two main points: (1) that SMR models are often biased by inflated estimates of significance (i.e., Type I errors); and (2) simpler models with fewer predictors are usually sufficient. We caution environmental biologists to recognise these limitations when using and interpreting SMR models. We also suggest that randomization-based probabilities are more reliable than traditional tabulated probabilities for assessing the results of SMR analyses.

Table of Contents

Executive Summary ii

Table of Contents iii
List of Tables iv
List of Figures v

Introduction 1

Randomization Algorithms 3

Data Sets
Simulated Data 4
Environmental Data 4

Simulation Results 5

Environmental Data Results 8

Discussion 13

Conclusions 14

Acknowledgements 14

References 14

List of Tables

Table 1: Randomized and tabulated probabilities for regression and ANOVA statistics from Ryder (1965) fish yield data set. The model predicts ln(fish yield) as a function of ln(Lake Area), ln(Lake Volume) and ln(TDS [or total dissolved solids]). 9

Table 2: Randomized and tabulated probabilities for regression and ANOVA statistics from Paloheimo and Zimmerman (1983) lake nutrient data set. The model predicts chlorophyll concentration as a function of total phosphorus, sediment area, calcium concentration, and the lake volume:sediment area ratio. 10

Table 3: Randomized and tabulated probabilities for regression and ANOVA statistics from the Rowan and Rasmussen (1992) data set predicting fish DDT levels from fisheries parameters, limnological variables, and environmental contaminant levels. 11

Table 4: Randomized and tabulated probabilities for regression and ANOVA statistics generated from the Rowan and Rasmussen (1992) data to predict fish PCB levels as a function of fisheries parameters, limnological variables, and environmental contaminant levels. 11

Table 5: Randomization and tabulated probabilities for regression and ANOVA statistics based on the model predicting angler fish abundance from environmental and oceanographic variables. 12

Table 6: Randomization and tabulated probabilities for regression and ANOVA statistics based on the model predicting angler fish abundance as a function of the scores derived from a Principal Components Analysis (PCA) of the matrix of fisheries and environmental parameters in Stergiou (1989). 12

List of Figures

Figure 1: Bias in tabulated probabilities for four different correlated predictor matrices. Data were simulated from Poisson, Gaussian, negative exponential, and rectangular (or uniform) distributions. All predictor matrices had inter-predictor correlations of 0.0. 5

Figure 2: Bias in tabulated probabilities for four different correlated predictor matrices. Data were simulated from Poisson, Gaussian, negative exponential, and rectangular (or uniform) distributions. All predictor matrices had inter-predictor correlations of 0.9. 6

Figure 3: Bias between randomized and tabulated probabilities for different levels of between predictor correlations and an N:P ratio of 20:5. 7

Figure 4: Difference between randomized and tabulated probabilities for varying levels of inter-predictor correlation assessed at an N:P ratio of 50:39. 7

Figure 5: Effect of N:P ratios on bias in tabulated probabilities. All tabulated probabilities were significant at P=0.05. Note that the higher the N:P ratio, the higher the bias. 8

Introduction

Meeting the competing demands of model parsimony and precision is a task faced by all environmental modellers. Stepwise multiple regression (SMR) is commonly used to build models that satisfy these competing criteria. SMR is a parametric modelling tool for predicting a response variable as a function of a subset of one or more parameters selected from a matrix of potential predictor variables. The method is designed to produce as good a compromise as possible between parsimony and precision (Altham 1984). Since the model produced is exclusively data driven, no account is taken of underlying theory, or prior judgement as to which parameters should be important in predicting a response. Due to its empirical nature, SMR analysis is popular in areas where substantive theory is weak.

Model building when substantive theory is weak is risky at best (Lovell 1983, Peters 1991). For example, significance estimates are artificially inflated when models are built from subsets of large numbers of predictor variables (Downing 1991). Regardless, SMR is often used in the environmental sciences, especially in exploratory studies where empirical modelling is used to generate new hypotheses (sensu Hakason and Peters 1995).

In this paper, we illustrate how randomization methods can be used to provide an unbiased estimate of the significance of empirical SMR models derived from environmental data sets. We also assess the consequences of violating assumptions underlying SMR. Randomization methods are used to address two basic questions: (1) what is the likelihood that the observed SMR model could occur due to chance alone if Y is truly random; and (2), how do the randomization-based significance levels differ from the standard tabulated values? Both of these questions are addressed by generating a population of F statistics derived from the observed and randomized data sets, and determining the number of randomized models which gave more extreme F statistics than reported for the observed model (e.g., Manly 1991).

SMR can be fit to data sets either through forward or backward variable selection procedures. Forward stepping models attempt to sequentially fit the predictor that explains a maximal amount of the variance in the response. The model evolves by adding other predictors that maximally explain the residual variation in the response variable. Backward stepping works in the reverse direction. A full model comprising all predictor variables is first fit to the data. Subsequently, predictors accounting for a non-significant amount of variation in the response variable are sequentially deleted. All analyses in this paper work with forward selection SMR models.

In a review of multivariate methods, James and McCulloch (1990) report a number of studies where SMR gave ambiguous results. A table of SMR models showed the selection of different sets of predictor variables depending on whether forward or backward stepping was employed. Similarly, bootstrapping the original data led to the selection of different subsets of predictors with SMR. Manly (1991, pp. 105-106) reported finding SMR model results that were significant at the 5% level in seven out of ten trials using random data for the response and predictors. The number of possible regression models that would have a tabulated significance value of $P<0.05$ due to chance alone depends on the number of potential predictors (Downing 1991). As the number of predictor variables increases, the

number of possible models increases exponentially. For P predictor variables, there are $2^P - 1$ possible regression models. With fifteen potential predictors, there are 32,767 possible models, of which 1638 will be significant at P=0.05 and 327 will be significant at P=0.01 due to chance alone. These findings cast doubt on the general utility of SMR as a tool for modelling complex environmental relationships, or for model building in the absence of underlying theory.

Rencher and Pun (1980) examined the problem of inflated R^2 values generated by SMR when the predictor variables are correlated, and when there are more predictor variables than observations. Their study was initiated after finding an SMR model with an R^2 of 0.98. By reshuffling the response variable such that it was randomly paired with the suite of predictors, they obtained another model with an R^2 of 0.95. In addition, suspiciously high R^2 values were noted for models built from uncorrelated predictor variables. The authors suggested that in cases where there are more variables than observations, "the results of stepwise regression may be of little value unless substantiated by another sample or other information independent of the data" (Rencher and Pun 1980). Frequently, there are more parameters than observations in environmental data sets. Using SMR as a model building technique in these instances should be avoided.

Numerous simulation studies (e.g., Freedman 1983, Flack and Chang 1987) have shown that uncorrelated, or random-noise variables are selected by SMR more often than would be expected due to chance alone. Unfortunately, these simulation studies are based on small sample sizes. Flack and Chang (1987) performed fifty simulations, and only two were performed by Freedman (1983). While these simulations suggested problems with SMR significance estimates, higher numbers of simulation runs would be more convincing. Improvements in computing power make it practical to perform much larger simulations than were possible even a few years ago.

Numerous problems have been reported with SMR in the biological (James and McCulloch 1990, Manly 1991, Downing 1991), statistical (Rencher and Pun 1980, Freedman 1983, Hoerl et al. 1983, Flack and Chang 1987, Roecker 1991) and econometric (Bacon 1977, Lovell 1983) literature. Despite a large body of evidence cautioning against its use, SMR is still commonly employed by environmental scientists. In this paper, we use computer intensive methods and randomization tests to demonstrate some statistical problems associated with SMR. Throughout this paper, we test the hypothesis that the response variable is unrelated to the predictor variables. Instead of using traditional parametric methods, we generate a population of test statistics for the hypothesis in question by randomizing or recombining the observed data (e.g., Manly 1991, ter Braak 1992). Thus, departure of the observed model from randomness can be assessed without fulfilling the usual assumptions associated with the use of statistical tables.

Like all parametric methods, SMR requires a number of assumptions about the data. These include normality, lack of collinearity among the predictors, and a sufficiently high ratio of observations to parameters. Failure to meet these assumptions can lead to inflated significance estimates (e.g., Somers 1997). The potential for these Type I errors can be further compounded when researchers transform or combine variables as ratios or products and enter both the transformed and untransformed variables into SMR (Downing 1991). These

manipulations will lower the ratio of observations to parameters, affect normality, and potentially increase the collinearity amongst the predictors.

There are two components to data normality: distributional assumptions, and the presence of outliers. It is assumed that regression models are derived from normally distributed data (Draper and Smith 1966, Fox 1991). These models can be extremely sensitive to outlying values (Fox 1991). Sensitivity to outliers has been discussed elsewhere (e.g., Fox 1991). Herein, the effects of violating assumptions about the underlying distribution from which the data were drawn will be assessed.

SMR models are sensitive to multi-collinearity (Stewart 1987). If variables in the predictor matrix are highly correlated with one another, problems with predictor selection (James and McCulloch 1990) and inflation of parameter standard error estimates (Fox 1991) may occur. It has been suggested that spurious relationships can be ruled out if inter-predictor correlations are less than 0.6 and all potential predictors are included in the analysis (Koslow et al. 1987).

Green (1979, p. 81) suggests that one solution to the multi-collinearity problem is to perform a PCA on the set of predictor variables and use the resultant, uncorrelated principal component scores in a regression model. Rencher and Pun (1980) observed that the SMR R^2 was inflated to the greatest degree with uncorrelated predictor variables, and thus, the likelihood of finding 'statistically significant' SMR models would be increased with the use of principal components.

To fit a regression model to a data set, there must be at least one more observation (N) than there are predictors (P). There are an infinite number of models available that describe a data set in which there are more parameters than observations (P>N). Unfortunately, P>N data sets are all too common in the environmental literature. Ideally, there should be at least ten observations per parameter before a regression model is fit to a data set (N:P=10). However, it has been suggested that a ratio as low as N:P=5 may be acceptable (Norman and Streiner 1986, p. 63).

Randomization Algorithms

To determine the likelihood that an observed SMR model could occur due to chance alone when the response variable is truly random, we implemented Draper and Smith's tableau algorithm for performing stepwise regressions (Draper and Smith 1966, pp. 177-193). The algorithm was modified using routines from Press et al. (1989) to calculate the appropriate F statistic based on the model degrees of freedom as opposed to the hard-coded F-statistic in the original algorithm. The algorithm for randomizing a response vector was adapted from Manly (1991). All code was written in C and executed on an HP9000 UNIX minicomputer.

Randomization-based estimates of significance were generated as follows (Manly 1991). The observed data set was run through the stepwise algorithm, and the model F statistic recorded. The randomization procedure was then iterated 1000 times. The randomization consisted of re-shuffling the response vector and re-entering it into the SMR analysis. The resultant F statistic for the re-shuffled response was recorded. In the event that the SMR algorithm was unable to fit a model to the data set, an F statistic of 0 was recorded for that run. By re-shuffling the response

variable and rerunning the SMR model, we produced a population of F statistics generated for each data set where the response variable was random with respect to the predictors. Significance of the original model was assessed by the rank order of the original F statistic in the distribution of randomized F statistics.

To determine the likelihood that the predictor variables selected in an SMR model actually contributed to the prediction of the response variable, we ran a series of randomizations using the parameters included in the observed SMR model. We re-shuffled the Y variable, recalculated the sums of squares and associated regression coefficients, and counted the number of sums of squares derived from the re-shuffled data that equalled or exceeded the observed sums of squares. The resultant probabilities for each predictor variable were used to determine whether or not each parameter in the observed model was a significant predictor of the response variable.

Data Sets

Simulated Data

Simulated data were used to evaluate the randomization approach to SMR when the response variable was truly random with respect to the predictor matrix. In all simulated data sets the predictor variables were uncorrelated with the response variable. Each simulated data set had similar correlations between all pairs of predictors. The average between-predictor correlation for each data set ranged from 0.0 to 0.99. The simulated data matrices were created in the following manner:

A (P+1) by (P+1) correlation matrix, **C**, was constructed to describe the desired correlations between the response and P predictor variables. An eigenanalysis of the correlation matrix, **C**, was performed. Eigenvalues, **e**, and eigenvectors, **V**, were retained. Next, a (P+1) by N matrix, **R**, was created and populated with random samples from a previously specified distribution (e.g., normal). The i^{th} column of R was multiplied by the square root of the i^{th} eigenvalue. This matrix was then multiplied by the matrix of eigenvectors, **V**, resulting in a (P+1) by N matrix, **A**, with the correlation structure originally specified in matrix **C**. This matrix **A** was used in the simulation analysis.

The effect of violating the assumption of data normality on SMR was examined by generating different predictor matrices from Gaussian, rectangular (uniform), Poisson, and negative exponential distributions. An algorithm from Press et al. (1989) was used to generate each of these distributions producing simulated data matrices of N:P = 20:10 with between-predictor correlations of 0.00 and 0.90.

To examine the effect of between-predictor collinearity on the randomization based SMR, we simulated matrices with uniform between-predictor correlations of 0.00, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.99. Each predictor variable was normally distributed with $\mu=0$ and $\sigma^2=1$.

The effect of varying the N:P ratio was examined by simulating matrices with N = 20, 30, 50 and 100 observations. For each number of observations, ten equally spaced sets of predictors were created to provide a range in N:P from N:P = N to N:P = 1. Each predictor variable was normally distributed with $\mu=0$ and $\sigma^2=1$.

Environmental Data

Randomization-based SMRs were evaluated using both real and simulated data. Four published environmental data sets were examined. These data sets contained information on fish yields from North American temperate lakes (Ryder 1965, Rempel and Colby 1991), contaminant levels in fish from the Laurentian Great Lakes (Rowan and Rasmussen 1992), lake nutrient-chlorophyll relationships (Paloheimo and Zimmerman 1983, Zimmerman et al. 1983), and factors influencing fish abundance (Stergiou 1989). These data sets were arbitrarily selected to represent the diversity of SMR applications in environmental sciences.

The Ryder (1965) data set was used to predict fish yield from lake area, lake volume and total dissolved solids (TDS). The Rowan and Rasmussen (1992) data set was used in SMR to predict organochlorine concentrations in fish from the Great Lakes as a function of biological and environmental variables. Biological variables included fish trophic level, length, weight, and lipid content. Environmental variables included sediment contaminant levels, lake depth, sedimentation rate, primary productivity, and secchi depth. The lake nutrient-chlorophyll data set was extracted from a regional lake survey (Paloheimo and Zimmerman 1983, Zimmerman et al. 1983). From the available data, we used chemical and physiographic variables (nitrogen, phosphorus, nitrogen:phosphorus ratios, lake area, lake volume, and mean depth) to predict summer chlorophyll concentrations.

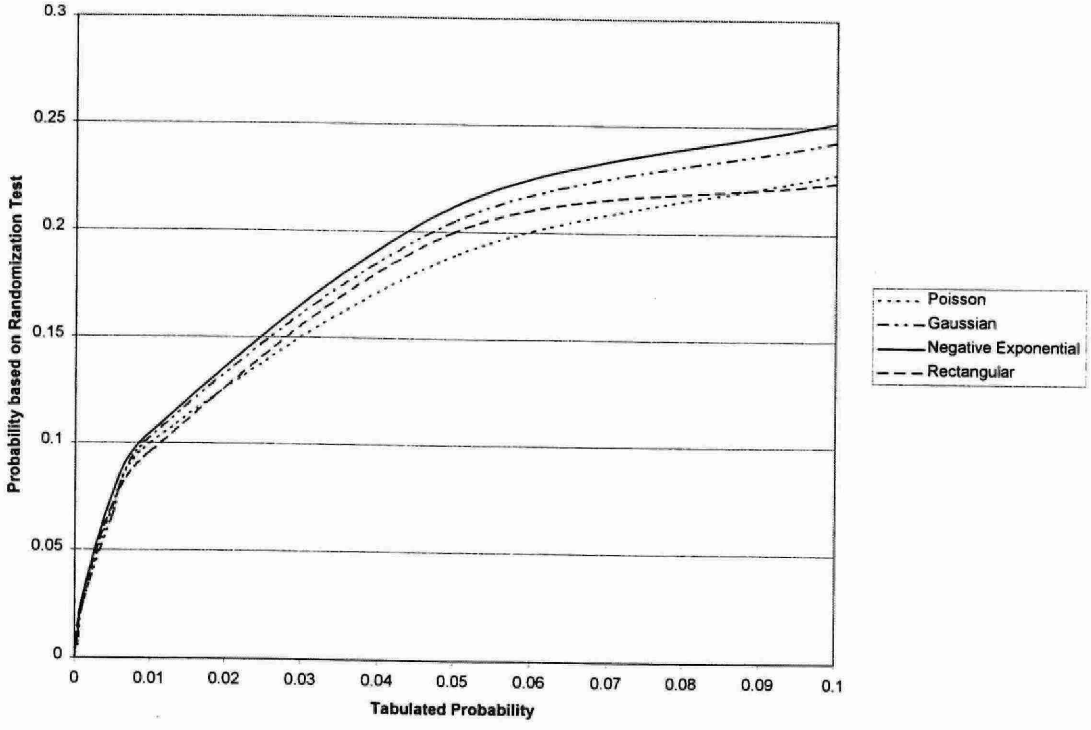


Figure 1: Bias in tabulated probabilities for four different correlated predictor matrices. Data were simulated from Poisson, Gaussian, negative exponential, and rectangular (or uniform) distributions. All predictor matrices had inter-predictor correlations of 0.0.

Simulation Results

Tabulated probabilities were biased for both the uncorrelated predictor matrix (Figure 1) and the multi-collinear predictors (Figure 2). However, the bias in the tabulated probability was relatively insensitive to the underlying distribution from which the data were drawn. Results were similar for models drawn from Gaussian, negative exponential, rectangular, and Poisson distributions. The F distribution was biased, although there was greater bias in the F distributions from the uncorrelated predictor matrix (Figure 1).

Given the results in Figures 1 and 2, greater bias in tabulated probability was observed for models built from low between-predictor correlation matrices (Figures 3 and 4). The bias all but disappeared as the collinearity approached 1 (which approaches a simple regression with only one predictor). While the curves in Figures 3 and 4 both have a similar shape, the differences in randomized probability should be noted. That is, the randomized probabilities are much smaller in Figure 3, illustrating the results of analyses using an N:P = 20:5 matrix than they are for Figure 4 whereN:P=50:39.

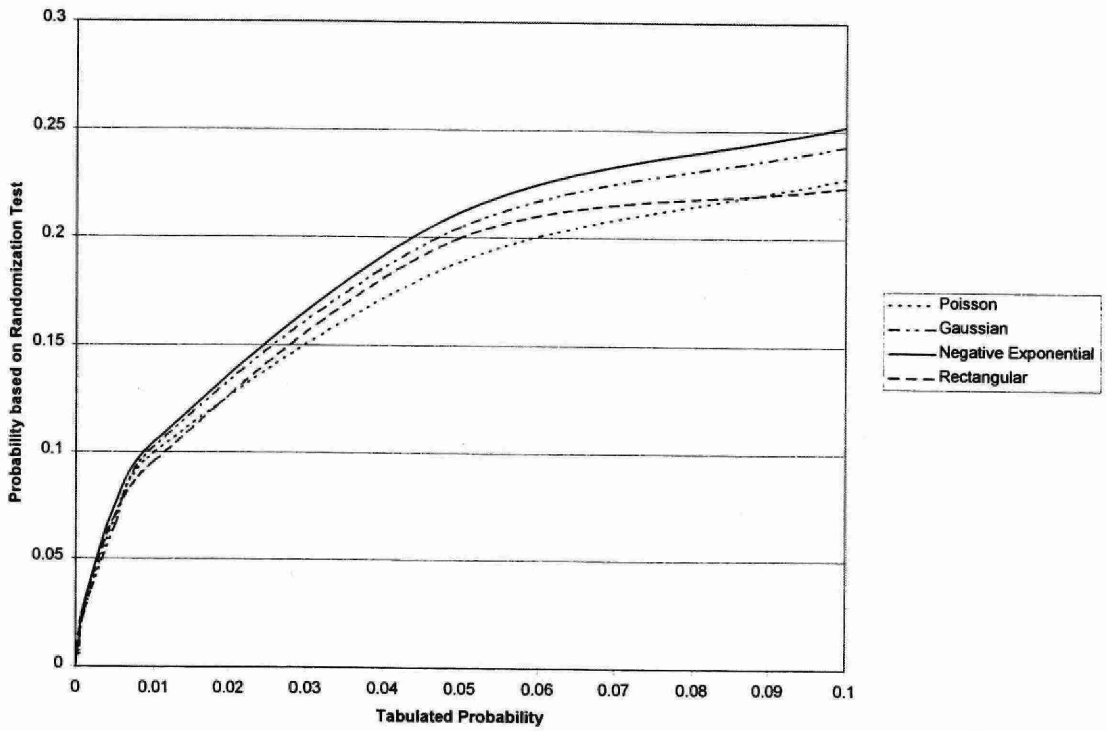


Figure 2: Bias in tabulated probabilities for four different correlated predictor matrices. Data were simulated from Poisson, Gaussian, negative exponential, and rectangular (or uniform) distributions. All predictor matrices had inter-predictor correlations of 0.9.

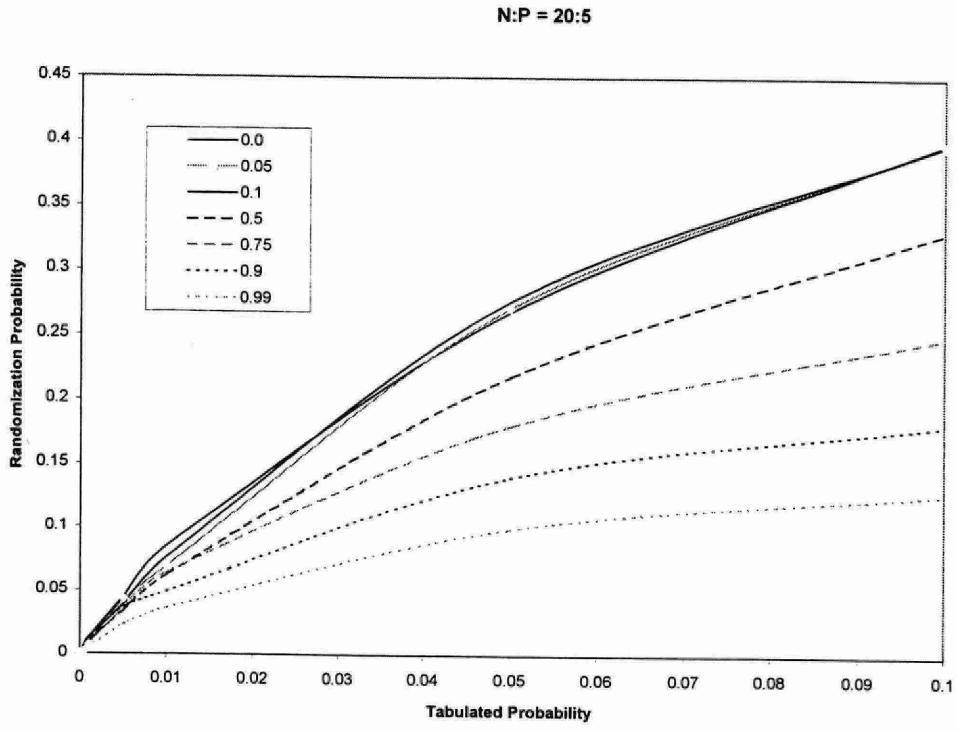


Figure 3: Bias between randomized and tabulated probabilities for different levels of between predictor correlations and an N:P ratio of 20:5.

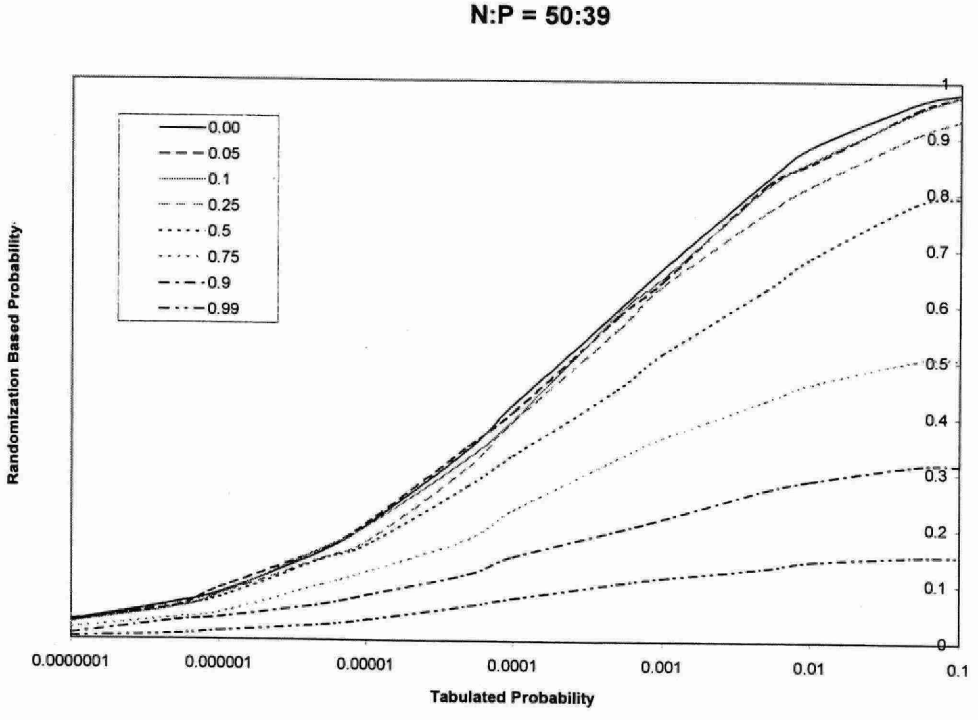


Figure 4: Difference between randomized and tabulated probabilities for varying levels of inter-predictor correlation assessed at an N:P ratio of 50:39.

It can be seen from Figure 5 that when only a single predictor variable was employed (N:P=N), there was no bias between tabulated and randomization based probabilities. As the number of predictors increased, so did the bias. This effect is compounded by the inter-predictor correlation since the bias is much worse for uncorrelated predictor matrices.

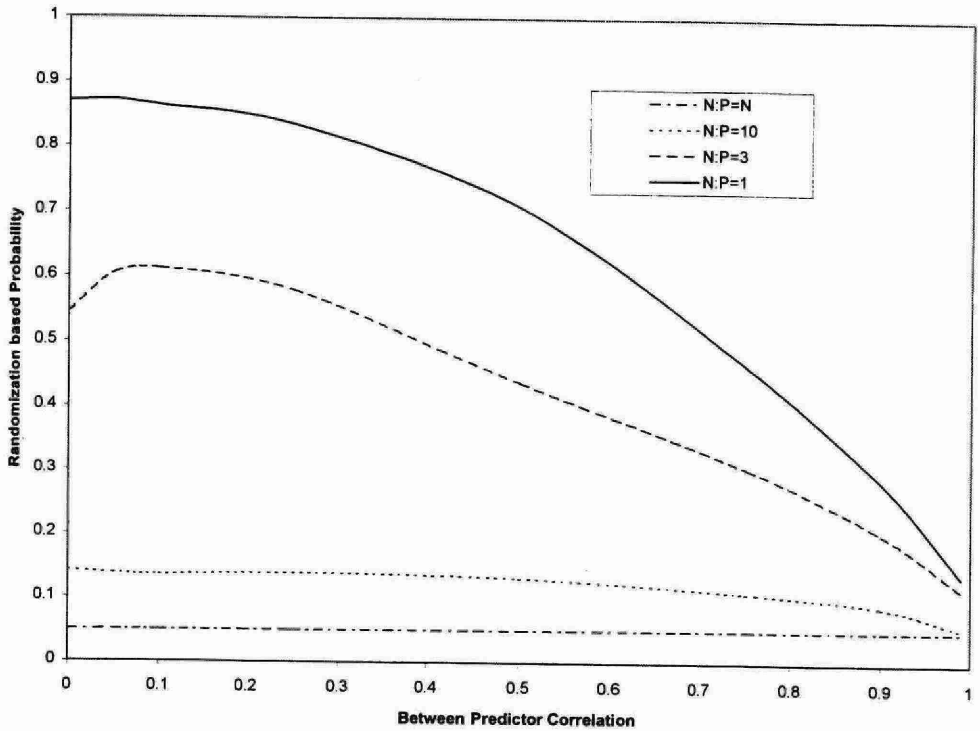


Figure 5: Effect of N:P ratios on bias in tabulated probabilities. All tabulated probabilities were significant at P=0.05. Note that the higher the N:P ratio, the higher the bias.

Environmental Data Results

For each environmental data set we report the following: (i) the observed SMR model, (ii) the ANOVA table for the observed model, (iii) the probability that each parameter contributed significantly to the final model, and (iv) the probability of encountering a model with a higher F statistic based on a randomization test. In addition, the tabulated and randomization-based significance estimates (P values) are presented.

Using Ryder's (1965) data on fish yields from 23 lakes, an SMR model with significant

tabulated and randomization-based probabilities resulted (P<0.001, Table 1). The model predicted fish yield as a function of lake volume, lake area, and concentration of total dissolved solids (TDS) in the lake. Lake volume was the strongest predictor of fish yield. By contrast, lake area was marginally significant (P=0.041) whereas TDS was not significant (P=0.218). Interestingly, the randomization test suggested that fish yield was not significantly predicted by TDS after the effects of lake area and lake volume were incorporated.

Variable		Coefficient	Partial R^2	P-values	
				Tabulated	Randomized
Intercept		1.727			
ln(Volume)		-0.561	0.512	<0.001	0.001
ln(Area)		0.588	0.175	0.001	0.041
ln(TDS)		0.294	0.074	0.024	0.218
ANOVA					
Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	7.541	3	20.354	<0.001	0.001
Residual	2.326	19			

Table 1: Randomized and tabulated probabilities for regression and ANOVA statistics from the Ryder (1965) fish yield data set. The model predicts ln(fish yield) as a function of ln(Lake Area), ln(Lake Volume) and ln(TDS [or total dissolved solids]).

Similarly, an SMR model was built to predict lake chlorophyll concentration as a function of lake morphometry and chemistry using the Paloheimo and Zimmerman (1983) and Zimmerman et al. (1983) data sets (Table 2). A highly significant model with randomized and tabulated probabilities of ~0.001 was generated using lake phosphorus concentration, sediment area, calcium concentration, and the lake volume to sediment area ratio. Further examination of the model coefficients revealed that only phosphorus and sediment area contributed substantially to variation in lake chlorophyll concentration (e.g., see the partial R^2 values). The predictive ability of calcium and the lake volume to sediment area ratio to explain variability in lake chlorophyll was not significantly different from what would be expected with random variables (P>0.25).

Variable	Coefficient	Partial R^2	P-values	
			Tabulated	Randomized
Intercept	-0.676			
Total Phosphorus	-0.332	0.512	<0.001	0.001
Sediment Area	0.491	0.283	<0.001	0.026
Calcium	0.034	0.011	0.014	0.276
Volume:Sediment Area	0.004	0.054	0.059	0.522

ANOVA Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	45.563	4	29.024	<0.001	0.001
Residual	7.457	19			

Table 2: Randomized and tabulated probabilities for regression and ANOVA statistics from the Paloheimo and Zimmerman (1983) lake nutrient data set. The model predicts chlorophyll concentration as a function of total phosphorus, sediment area, calcium concentration, and lake volume:sediment area ratio.

A data set on organic contaminants in fish from the Laurentian Great Lakes was obtained from Rowan and Rasmussen (1992). Randomization based estimates of model probabilities yielded interesting results (Tables 3 and 4). Highly significant (P=0.001) tabulated and randomized probabilities were obtained for both models. Concentrations of DDT and PCB in fish were significantly predicted by fish lipid content, the trophic level of the fish (coded as either 0 for a planktivore or 1 for a piscivore), and the ratio of fish yield to primary production. The standard model predicting DDT in fish (Table 3) erroneously included the concentration of DDT in sediments and suspended sediment concentration (P>0.10). Unlike the DDT model, the SMR predicting PCB concentration in fish produced similar results for both the standard and randomisation-based models (Table 4). This finding suggests that the PCB data set fulfilled all of the parametric assumptions underlying traditional SMR.

Variable	Coefficient	Partial R^2	P-values	
			Tabulated	Randomized
Intercept	2.196			
$\log_{10}$(FY/PP)	-0.267	0.279	<0.001	0.001
Trophic Level	0.329	0.087	<0.001	0.002
$\log_{10}$(Lipid)	0.502	0.114	<0.001	0.002
$\log_{10}$(Sediment DDT)	0.505	0.026	<0.001	0.141
$\log_{10}$(Sedimentation Rate)	-0.325	0.085	0.003	0.007
$\log_{10}$(Suspended Sediment)	-0.205	0.018	0.047	0.199

ANOVA Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	12.834	6	22.825	<0.001	0.001
Residual	8.247	88			

Table 3: Randomized and tabulated probabilities for regression and ANOVA statistics from the Rowan and Rasmussen (1992) data set predicting fish DDT levels from fisheries parameters, limnological variables, and environmental contaminant levels.

Variable	Coefficient	Partial R^2	P-values	
			Tabulated	Randomized
Intercept	1.753			
$\log_{10}$(Lipid)	0.692	0.247	<0.001	0.001
$\log_{10}$(Water PCB)	1.171	0.175	<0.001	0.001
$\log_{10}$(FY/PP)	-0.404	0.201	<0.001	0.001
Trophic Level	0.336	0.115	<0.001	0.001

ANOVA Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	16.603	4	71.777	<0.001	0.001
Residual	5.899	102			

Table 4: Randomized and tabulated probabilities for regression and ANOVA statistics generated from the Rowan and Rasmussen (1992) data to predict fish PCB levels as a function of fisheries parameters, limnological variables and environmental contaminant levels.

Two models were built predicting angler fish abundance. One model used the observed predictor matrix (Table 5) whereas the second SMR used the principal component scores (PCs) generated from a PCA of the original environmental predictors (Table 6). A significant model was constructed predicting angler fish abundance as a function of both temperature and prey abundance (Table 5). However, randomization tests indicated that only temperature contributed significantly to the model. Interestingly, the probability of finding more significant models due to chance alone is higher for the analysis based on the PCs than for the raw variable SMR (P=0.028 vs. P=0.018). This result is consistent with our simulations that showed greater bias in regression F statistics from SMR models using uncorrelated predictor variables.

Variable	Coefficient	Partial R^2	P-values	
			Tabulated	Randomized
Intercept	40.297			
$\log_{10}$(Temperature)	-14.979	0.451	0.019	0.006
$\log_{10}$(Prey Abundance)	0.255	0.111	0.082	0.240

ANOVA Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	10.107	2	8.952	0.004	0.018
Residual	7.903	14			

Table 5: Randomization and tabulated probabilities for regression and ANOVA statistics based on the model predicting angler fish abundance from environmental and oceanographic variables.

Variable	Coefficient	Partial R^2	P-values	
			Tabulated	Randomized
Intercept	1.220			
PC 1	-0.622	0.390	0.004	0.009
PC 3	0.411	0.150	0.051	0.117

ANOVA Source	Sum of Squares	df	F-ratio	P-values	
				Tabulated	Randomized
Regression	9.720	2	8.209	0.005	0.028
Residual	8.289	14			

Table 6: Randomization and tabulated probabilities for regression and ANOVA statistics based on the model predicting angler fish abundance as a function of the scores derived from a Principal Components Analysis (PCA) of the fisheries and environmental parameters in Stergiou (1989).

In general, the randomization results from our analysis of the environmental data sets support our findings obtained with simulated data: models with uncorrelated predictor variables are more subject to bias in SMR than are those models produced with correlated predictor variables. This finding is consistent with the concern that SMR models based on small N:P ratios (i.e., with more predictors) have higher Type I error rates (e.g., Downing 1991). In addition, randomization-based models with R^2 values similar to the observed models can often be obtained using fewer predictors than traditional SMR models using tabulated probabilities.

Discussion

Analysis of simulated data sets demonstrated the usefulness of the randomization approach in quantifying the consequences of violating three SMR assumptions. First, the data must be drawn from a normal distribution; second, there must be a sufficiently high ratio of observations to variables; and third, the predictor variables must not be collinear. In particular, our randomization tests showed that violating the latter two assumptions will lead to erroneously high estimates of model significance (i.e., Type I errors indicating significance when they shouldn't).

It has been suggested that as long as the average inter-predictor correlation is less than 0.6, it will be possible to identify spurious models (Koslow et al. 1987). Our results led to a different conclusion. That is, the likelihood of encountering bias in the regression F statistics increases as the inter-predictor correlation decreases. For highly inter-correlated predictor variables, there is, in effect, a single signal among the predictor variables to which differing amounts of random noise have been added. However, for a matrix of P uncorrelated predictor variables, there are P independent signals and there are $(2^P - 1)$ possible combinations of the P predictors (Downing 1991). Given this large number of combinations, it should be apparent that $0.05 \times (2^P - 1)$ of these combinations will be significant at the P=0.05 level by chance alone. We believe that the randomization-based probabilities were superior to the tabulated probabilities when multi-collinearity (and conversely, the N:P ratio) was an issue.

Data-reduction procedures such as PCA have been advocated as a solution to the multi-collinearity problem in SMR (e.g., Green 1979, Stergiou 1991). While PC scores obviously are not multi-collinear, our results show that using uncorrelated predictors, such as PC scores, can increase the likelihood that SMR will find a statistically significant model. As noted above, randomization-based probabilities appear to be robust to these sorts of Type I errors.

Lastly, we believe that our results demonstrate that significant SMR models can be readily found when the response variable is unrelated to the predictors. That is, we generated an unexpectedly large number of SMR models with tabulated significance values of $P<0.05$

from randomized environmental data sets. Frequently, we found tabulated significance estimates that were orders of magnitude smaller than the randomisation-based probabilities. As a result, we recommend caution when assessing the statistical significance of SMR models based on environmental data and traditional tabulated probabilities.

Although we examined a number of issues in SMR in this study, many issues remain: Does the backwards elimination, step-down algorithm lead to results that are similar to those presented here?; What is the likelihood that single-predictor models will be produced from random data?; and, once such a model has been produced, Will the forward stepwise algorithm add another predictor? Results of these and future assessments should have bearing on the interpretation of stepwise procedures in other methods such as discriminant analysis and canonical correspondence analysis (e.g., Rencher and Larson 1980, ter Braak 1986).

Conclusions

Our results emphasize two main points: (1) that SMR models are often biased by inflated estimates of significance (i.e., Type I errors); and (2), simpler models with fewer predictors are usually sufficient. We caution environmental biologists to recognise these limitations when using and interpreting SMR models. We also recommend that randomization-based probabilities be used to evaluate SMR models resulting from environmental data.

Acknowledgements

We thank Peter Dillon, Don Jackson, and Roger Green for comments on earlier versions of this report. Discussions with Don Jackson and Bryan Manly helped to focus our objectives and ultimately, the questions addressed herein.

References

Altham, P.M.E., 'Improving the precision of estimation by fitting a model', *Journal of the Royal Statistical Society, Series B*, **46,** 118-119 (1984).

Bacon, R.W., 'Some evidence on the squared correlation coefficient from several samples', *Econometrica*, **45,** 1997-2001 (1977).

Downing, J.A., 'Comparing apples with oranges: methods of inter-ecosystem comparison' in Cole, J., Lovett, G. and Findlay, S. (eds.), *Comparative Analysis of Ecosystems*, Springer-Verlag, Berlin, 1991, pp. 24-45.

Draper, N.R. and Smith, H., *Applied Regression Analysis*, John Wiley & Sons, Inc. New York, 1966, 407 p.

Flack, V.F. and Chang, P.C., 'Frequency of selecting noise variables in subset regression analysis: a simulation study', *The American Statistician*, **41,** 84-86 (1987).

Fox, J., *Regression diagnostics*, Sage Publications, Beverly Hills, CA, 1991, 96 p.

Freedman, D.A., 'A note on screening regression equations', *The American Statistician*, **37,** 152-155 (1983).

Green, R.H., *Sampling design and statistical methods for environmental biologists*, John Wiley & Sons Inc, New York, 1979, 257 p.

Hakason, L. and R.H. Peters. 1995, *Predictive Limnology - methods for predictive modelling*, SPB Academic Publishing, Amsterdam, 464 p.

Hoerl, R.W., Schuenmayer, J.H. and Hoerl, A.E., 'A simulation of biased estimation and subset selection regression techniques', *Technometrics*, **28,** 369-380 (1986).

James, F.C. and McCulloch, C.E., 'Multivariate analysis in ecology and systematics: panacea or Pandora's box' in *Annual Review of Ecology and Systematics,* Annual Reviews Inc., Palo Alto, CA. 1990, pp. 129-166.

Koslow, J.A., Thompson, K.R. and Silvert, W., 'Recruitment of northwest Atlantic Cod (*Gadus morhua*) and Haddock (*Melanogrammus aeglefinus*) stocks: influence of stock size and climate', *Can. J. Fish. Aquat. Sci.*, **44,** 26-39 (1987).

Lovell, M.C., 'Data mining', *The Review of Economics and Statistics*, **65,** 1-12 (1983).

Manly, B.F.J., *Randomization and Monte Carlo Methods in Biology*, Chapman and Hall, London, 1991, 281 p.

Norman, G.R. and Streiner, D.L.,*PDQ Statistics*, B.C. Decker, Inc., Toronto, 1986, 172 p.

Paloheimo, J.E. and Zimmerman, A.P., 'Factors influencing phosphorus-phytoplankton relationships', *Can. J. Fish. Aquat. Sci.*, **40,** 1804-1812 (1983).

Peters, R.H., *A critique for ecology*, Cambridge University Press, Cambridge, 1991, 384 p.

Press, W.H., Flannery B.P., Teukolsky, S.A. and Vetterling, W.T. *Numerical Recipes in C*, Cambridge University Press, Cambridge, 1989, 735 p.

Rempel, R.S. and Colby, P.C., 'A statistically valid model of the morphoedaphic index', *Can. J. Fish. Aquat. Sci.*, **48,** 1937-1943 (1991).

Rencher, A.C. and Larson, S.F., 'Bias of Wilks' Λ in stepwise discriminant analysis', *Technometrics*, **22,** 349-356 (1980).

Rencher, A.C. and Pun, F.C., 'Inflation of R^2 in best subset regression', *Technometrics*, **22,** 49-53 (1987).

Roecker, E.B., 'Prediction error and its estimation for subset-selected models', *Technometrics*, **33,** 459-468 (1991).

Rowan, D.J. and Rasmussen, J.B., 'Why don't Great Lakes fish reflect environmental concentrations of organic contaminants? - an analysis of between-lake variability in the ecological partitioning of PCBs and DDT', *Journal of Great Lakes Research*, **18,** 724-741 (1992).

Ryder, R.A., 'A method for estimating the potential fish production of north-temperate lakes', *Transactions of the American Fisheries Society*, **94,** 214-218 (1965).

Somers, K.M. 1997. Power Analysis: A statistical tool for assessing the utility of a study. Dorset Environmental Science Centre, Ontario Ministry of the Environment, Dorset, Ontario. ISBN: 0-7778-6958-6.

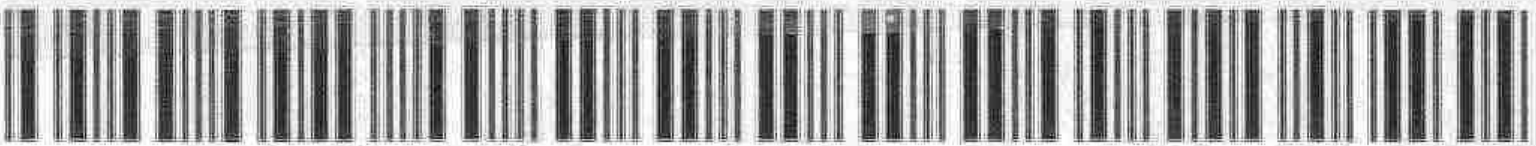
Stergiou, K.I., 'A method to cope with collinearity of ecological data sets in community studies', *Coenoses*, **4,** 91-94 (1989).

Stewart, G.W., 'Collinearity and least squares regression', *Statistical Science*, **2,** 68-84 (1987).

ter Braak, C.J.F., 'Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis'. *Ecology*, **67**, 1167-1179 (1986).

ter Braak, C.J.F., 'Permutation versus bootstrap significance tests in multiple regression and ANOVA' in *Bootstrapping and Related Techniques*, Springer Verlag, Berlin, 1992, pp.79-86.

Zimmerman, A.P., Noble, K.M., Gates, M.A. and Paloheimo, J.E., 'Physicochemical typologies of south-central Ontario lakes', *Can. J. Fish. Aquat. Sci.*, **40,** 1788-1803 (1983).


9693600000 9427